



SRCC Workshop on Advanced Statistical and Machine Learning Methods

Date : 29 April 2024 (Monday)

Time : 14:00–17:30

Venue : FSC1217, Fong Shu Chuen Library
Ho Sin Hang Campus
Hong Kong Baptist University

Xu GUO (School of Statistics, Beijing Normal University), **14:00-14:40**

Title: Model-free Variable Importance Detection with Machine Learning Methods

Abstract: In this paper, we investigate variable importance testing problem in a model-free framework. Some remarkable procedures are developed recently. Despite their success, existing procedures suffer from a significant limitation, that is, they generally require larger training sample and do not have the fastest possible convergence rate under alternative hypothesis. In this paper, we propose a new procedure to test variable importance. Flexible machine learning methods are adopted to estimate unknown functions. Under null hypothesis, our proposed test statistic converges to standard chi-squared distribution. While under local alternative hypotheses, it converges to non-central chi-square distribution. It has non-trivial power against the local alternative hypothesis which converges to the null at the fastest possible rate. We also extend our procedure to test conditional independence. Asymptotic properties are also developed. Numerical studies and two real data examples are conducted to illustrate the performance of our proposed test statistic.

Baihua HE (Department of Statistics and Finance, University of Science and Technology of China), **14:40-15:20**

Title: Deep Representation Transfer Learning for Partially Linear Models

Abstract: Transfer learning has achieved many successes in practical applications. While some certain issues are remaining unexplored, especially the combination of parameter interpretability and model flexibility. In this article, we present a novel deep neural network-based transfer learning approach within a partially linear model to enhance model performance using heterogeneous source domain data. The proposal addresses the challenges of high-dimensional and non-linear data, offering both high prediction accuracy and improved interpretability of primary parameters. Distinct from existing statistical methods, our framework facilitates positive transfer under certain domain heterogeneity, maintaining robustness and efficiency through data augmentation. We establish the consistency of representation learning, asymptotic normality of the primary parameters as well as the semi-parametric efficiency. We demonstrate the promising performance of the proposal through simulations and empirical validations on house renting price study and reemployment bonus experiment.



Feifei CHEN (Faculty of Arts and Sciences, Beijing Normal University at Zhuhai), **15:20-16:00**

Title: Model Checks on Single-index Varying Coefficient Models with Functional Response

Abstract: When building a regression model, the interpretations could be misleading if the model is not adequate. It is therefore necessary to check the adequacy of a model. This article aims to propose a model checking method for single-index varying coefficient models with functional response. A residual-marked empirical process-based global test is proposed. The test is projection-based, which can well circumvent the curse of dimensionality. Simulation studies show the good performance of the proposed method in various scenarios. For illustration, we test the single-index varying coefficient model structure on a real data set. An R package is made available online to implement our proposed test.

Break Time: 16:00-16:10

Yan ZHOU (School of Mathematical Sciences, Shenzhen University), **16:10-16:50**

Title: GCM-DiffHiC: Gaussian Convolution-based Method for Differential Chromatin Interactions Detection

Abstract: High-throughput techniques bring novel tools and also statistical challenges to genomic research. Variations in the three-dimensional genome structure can affect understanding of genome expression and regulation. However, there are spatial dependencies between adjacent DNA fibers involved in chromatin loop formation, posing a challenge in testing differences in Hi-C data. In this paper, we present a new method called GCM-DiffHiC to detect different chromatin interaction. Considering the correlation between interactions, a Gaussian filter is used to weigh Hi-C data. Under the assumption that interaction readings obey Poisson distribution, a hypothesis testing framework is established. Using the available knowledge that there are no differences in Hi-C diagonal interactions, an optimal scaling factor is determined to minimize the deviation between empirical and nominal I-type errors. Interactions are then tested using the optimal scaling factor. Simulation studies and real data analysis show that our method is more accurate and robust than other alternatives for detecting differential chromatin interactions.

Xiaochen ZHANG (Faculty of Arts and Sciences, Beijing Normal University at Zhuhai), **16:50-17:30**

Title: Functional Linear Model with Prior Information of Subjects' Network

Abstract: In many modern applications, data samples are interconnected by a network, and network information is a crucial factor in forecasting. However, existing network data analysis methods, which are designed for scalar data, are not effective for infinite-dimensional function data, particularly when functional predictors are observed on an irregular sampling design. In this paper, we propose a functional linear model for network-linked data. To improve the estimation and prediction, the network cohesion is enforced using the Laplace quadratic penalty function. The statistical properties of the proposed model are studied, and an extension to high-dimensional functional data is developed to simultaneously select relevant functional predictors and estimate the coefficient functions. Simulation results and real data application demonstrate the satisfactory performance of the proposed methods.

-- All interested are most welcome! --